IN THE CLAIMS

Please cancel Claims 1-51.

Please add the following claims:

---

52. An apparatus for determining if a query document matches one or more documents in a database, the apparatus comprising:

means for identifying up endpoints and down endpoints in the query document, the up endpoints representing tops of features in the query document and the down endpoints representing bottoms of features in the query document;

means for generating a set of descriptors for the query document based on locations of the up endpoints and the down endpoints;

means for comparing the set of descriptors for the query document against respective sets of descriptors associated with the one or more documents in the database to determine if the query document matches at least one of the one or more documents;

wherein the means for generating a set of descriptors for the query document based on locations of the up endpoints and the down endpoints comprises

means for identifying text lines in the query document based on concentrations of up endpoints and down endpoints along scanlines of the query document; and

means for generating the set of descriptors based on distances between selected up endpoints and selected down endpoints within the text lines in the query document; and

wherein the means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

means for determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

means for identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

53. An apparatus for determining if a query document matches one or more documents in a database, the apparatus comprising:

means for generating a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the query document;

means for comparing the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to identify one or more candidate documents;

means for identifying endpoint features in the query document;

means for generating a set of descriptors for the query document based on locations of the endpoint features;

means for comparing the set of descriptors for the query document against respective sets of descriptors for the one or more candidate documents to determine if the query document matches at least one of the one or more candidate documents;

means for performing spectral analysis on the bit profile of the query document to determine global statistics of the query document; and

means for comparing the global statistics of the query document against global statistics associated with a second plurality of documents from the database to identify the first plurality of documents, the first plurality of documents being a subset of the second plurality of documents.

54. The apparatus of claim 53 wherein the means for performing spectral analysis on the bit profile to determine global statistics comprises means for generating an estimation of at least one of a dominant line spacing in the query document, a proportion of the query document that is text, a location of text in the query document, and a text concentration.

55. An apparatus for generating a set of descriptors for identifying a document, the apparatus comprising:

means for identifying up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints;

wherein the means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

means for determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

means for identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

56. An apparatus for generating a set of descriptors for identifying a document, the apparatus comprising:

means for identifying up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints;

wherein the means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

means for determining a dominant line spacing in the document;

means for determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

means for identifying as text lines respective scanline pairs in which the constituent scanlines are separated by a distance less than the dominant line spacing and in which the constituent scanlines respectively have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

57. The apparatus of claim 56 wherein the dominant line spacing is determined based on spectral analysis of locations of the endpoints in the document.

58. An apparatus for generating a set of descriptors for identifying a document, the apparatus comprising:

means for identifying up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document;

means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints; and

means for generating a respective endpoint profile for each of the scanlines, the endpoint profile including a count of up endpoints identified on the scanline and a count of down endpoints identified on the scanline, and wherein the means for identifying text lines based on concentrations of up endpoints and down endpoints along scanlines of the document comprises means for reducing all but local maximums of the counts of up endpoints and the counts of down endpoints in respective endpoint profiles.

59. An apparatus for generating a set of descriptors for identifying a document, the apparatus comprising:

means for identifying up endpoints and down endpoints in the

document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints;

wherein the means for identifying text lines based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

means for generating a count of up endpoints and a count of down endpoints for each of the scanlines;

means for identifying a first scanline within a locality of scanlines that has the highest count of up endpoints;

means for reducing the count of up endpoints associated with each scanline within the locality of scanlines except the first scanline;

means for identifying a second scanline within the locality of scanlines that has the highest count of down endpoints; and

means for reducing the count of down endpoints associated with each scanline within the locality of scanlines except the second scanline.

60. The apparatus of claim 59 wherein the means for identifying the first scanline within the locality of scanlines that has the highest count of up endpoints comprises:

means for determining a dominant line spacing of the document; and

means for defining the locality of scanlines to be scanlines within a range greater than the dominant line spacing but less than twice the dominant line

spacing.

61. An apparatus for generating a set of descriptors for identifying a document, the apparatus comprising:

means for identifying up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

means for identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints;

wherein the means for generating a set of descriptors based on distances between selected up endpoints and selected down endpoints comprises means for defining an ascender zone and a descender zone for each of the text lines, the selected up endpoints being up endpoints in the ascender zone and the selected down endpoints being down endpoints in the descender zone.

62. The apparatus of claim 61 wherein the means for defining an ascender zone and a descender zone for each of the text lines comprises:

means for defining a region above an x-height line of a first text line of the text lines to be the ascender zone for the first text line; and

means for defining a region below the baseline of the first text line to be the descender zone for the first text line.

63. The apparatus of claim 62 wherein the ascender zone of the first

text line is bounded in part by the descender zone for the preceding text line.

64. An apparatus for generating information that can be used to identify a document, the apparatus comprising:

means for generating a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

means for performing spectral analysis on the bit profile to determine global statistics of the document including means for generating an estimation of a dominant line spacing in the document, wherein the means for generating an estimation of a dominant line spacing comprises means for generating a power spectrum density from the bit profile and means for calculating the estimation of the dominant line spacing from a peak value in the power spectrum density.

65. An apparatus for generating information that can be used to identify a document, the apparatus comprising:

means for generating a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

means for performing spectral analysis on the bit profile to determine global statistics of the document, wherein the means for performing spectral analysis on the bit profile to determine global statistics comprises means for generating an estimation of a proportion of the document that is text, and further wherein the means for generating an estimation of a proportion of the document that is text comprises means for generating a power spectrum density from the bit profile and means for calculating the estimation of the proportion of the document based on an energy under a peak value in the power spectrum

density.

66. An apparatus for generating information that can be used to identify a document, the apparatus comprising:

means for generating a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document;

means for performing spectral analysis on the bit profile to determine global statistics of the document, wherein means for performing spectral analysis on the bit profile to determine global statistics comprises means for generating an estimation of a location of text in the document, and wherein the means for generating an estimation of a location of text in the document comprises

means for applying a bandpass filter to the bit profile to generate a text energy profile, and

means for determining a centroid of the text energy profile to be the estimation of the location of text in the document.

67. The apparatus of claim 66 wherein the means for applying a bandpass filter to the bit profile comprises:

means for determining a dominant line spacing frequency of the document; and

means for selecting a center frequency of the bandpass filter based on the dominant line spacing frequency.

68. An apparatus for generating information that can be used to identify a document, the apparatus comprising:

means for generating a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

means for performing spectral analysis on the bit profile to determine global statistics of the document, wherein the means for performing spectral analysis on the bit profile to determine global statistics comprises the means for generating an estimation of text concentration in the document, the estimation of text concentration indicating a lengthwise measure of a proportion of the document that is text, and further wherein the means for generating an estimation of text concentration in the document comprises:

means for applying a bandpass filter to the bit profile to generate a text energy profile; and

means for determining the estimation of the text concentration based on a length of the text energy profile.

69. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the query document, the up endpoints representing tops of features in the query document and the down endpoints representing bottoms of features in the query document;

generate a set of descriptors for the query document based on locations of the up endpoints and the down endpoints by

identifying text lines in the query document based on

concentrations of up endpoints and down endpoints along scanlines of the query document by,

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines; and

generating the set of descriptors based on distances between selected up endpoints and selected down endpoints within the text lines in the query document;

compare the set of descriptors for the query document against respective sets of descriptors associated with the one or more documents in the database to determine if the query document matches at least one of the one or more documents;

70. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

generate a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the query document;

compare the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to identify one or more candidate documents;

identify endpoint features in the query document;

generate a set of descriptors for the query document based on locations of

the endpoint features;

compare the set of descriptors for the query document against respective sets of descriptors for the one or more candidate documents to determine if the query document matches at least one of the one or more candidate documents;

perform spectral analysis on the bit profile of the query document to determine global statistics of the query document; and

compare the global statistics of the query document against global statistics associated with a second plurality of documents from the database to identify the first plurality of documents, the first plurality of documents being a subset of the second plurality of documents.

71. The method of claim 70 wherein the instructions to perform spectral analysis on the bit profile to determine global statistics comprises instructions, which when executed by the system cause the system to generate an estimation of at least one of a dominant line spacing in the query document, a proportion of the query document that is text, a location of text in the query document, and a text concentration.

72. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identify text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document by

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines, and

identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines; and

generate a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints.

73. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identify text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document by

determining a dominant line spacing in the document,

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines, and

identifying as text lines respective scanline pairs in which the constituent scanlines are separated by a distance less than the dominant line spacing and in which the constituent scanlines respectively have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines; and

generate a set of descriptors based on distances between selected up

endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints.

74. The method of claim 73 wherein the dominant line spacing is determined based on spectral analysis of locations of the endpoints in the document.

75. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identify text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document;

generate a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints; and

generate a respective endpoint profile for each of the scanlines, the endpoint profile including a count of up endpoints identified on the scanline and a count of down endpoints identified on the scanline, and wherein identifying text lines based on concentrations of up endpoints and down endpoints along scanlines of the document comprises reducing all but local maximums of the counts of up endpoints and the counts of down endpoints in respective endpoint profiles.

76. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identify text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document by

generating a count of up endpoints and a count of down endpoints for each of the scanlines,

identifying a first scanline within a locality of scanlines that has the highest count of up endpoints,

reducing the count of up endpoints associated with each scanline within the locality of scanlines except the first scanline,

identifying a second scanline within the locality of scanlines that has the highest count of down endpoints, and

reducing the count of down endpoints associated with each scanline within the locality of scanlines except the second scanline; and

generate a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints.

77. The method of claim 76 wherein instructions to identify the first scanline within the locality of scanlines that has the highest count of up endpoints comprise instructions which when executed by the system cause the

system to:

determine a dominant line spacing of the document; and

define the locality of scanlines to be scanlines within a range greater than the dominant line spacing but less than twice the dominant line spacing.

78. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

identify up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identify text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

generate a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints;

wherein the set of descriptors are generated by defining an ascender zone and a descender zone for each of the text lines, the selected up endpoints being up endpoints in the ascender zone and the selected down endpoints being down endpoints in the descender zone.

79. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

define a region above an x-height line of a first text line of the text lines to be the ascender zone for the first text line; and

define a region below the baseline of the first text line to be the descender zone for the first text line.

80. The method of claim 79 wherein the ascender zone of the first text line is bounded in part by the descender zone for the preceding text line.

81. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

generate a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

perform spectral analysis on the bit profile to determine global statistics of the document;

wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a dominant line spacing in the document; and

wherein generating an estimation of a dominant line spacing comprises generating a power spectrum density from the bit profile and calculating the estimation of the dominant line spacing from a peak value in the power spectrum density.

82. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

generate a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document by;

perform spectral analysis on the bit profile to determine global statistics of the document by generating an estimation of a proportion of the document that is text, wherein generating an estimation of a proportion of the document that is text comprises generating a power spectrum density from the bit profile and calculating the estimation of the proportion of the document based on an energy under a peak value in the power spectrum density.

83. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

generate a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

perform spectral analysis on the bit profile to determine global statistics of the document by generating an estimation of a location of text in the document, wherein generating an estimation of a location of text in the document comprises

applying a bandpass filter to the bit profile to generate a text energy profile; and

determining a centroid of the text energy profile to be the estimation of the location of text in the document.

84. The method of claim 83 wherein instructions for applying a bandpass filter to the bit profile comprises instructions which when executed by the system cause the system to:

determine a dominant line spacing frequency of the document; and

selecting a center frequency of the bandpass filter based on the dominant line spacing frequency.

85. An article of manufacture having one or more recordable media with executable instructions stored thereon which, when executed by a system, cause the system to:

A2 Cont'd

generate a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

perform spectral analysis on the bit profile to determine global statistics of the document by generating an estimation of text concentration in the document, the estimation of text concentration indicating a lengthwise measure of a proportion of the document that is text, wherein generating an estimation of text concentration in the document is performed by:

applying a bandpass filter to the bit profile to generate a text energy profile; and

determining the estimation of the text concentration based on a length of the text energy profile.